Normal Bandits of Unknown Means and Variances:
Asymptotic Optimality, Finite Horizon Regret Bounds, and a Solution to an Open Problem

Wesley Cowan CWCOWAN@MATH.RUTGERS.EDU
Department of Mathematics
Rutgers University
110 Frelinghuysen Rd., Piscataway, NJ 08854, USA

Junya Honda HONDA@IT.K.U-TOKYO.AC.JP
Department of Complexity Science and Engineering
Graduate School of Frontier Sciences, The University of Tokyo.
5-1-5 Kashiwanoha, Kashiwa-shi, Chiba 277-8561, Japan.

Michael N. Katehakis MNK@RUTGERS.EDU
Department of Management Science and Information Systems
Rutgers University
100 Rockafeller Rd., Piscataway, NJ 08854, USA

Abstract

Consider the problem of sampling sequentially from a finite number of $N \geqslant 2$ populations, specified by random variables X^i_k, $i = 1,\ldots,N$, and $k = 1,2,\ldots$; where X^i_k denotes the outcome from population i the k^{th} time it is sampled. It is assumed that for each fixed i, $\{X^i_k\}_{k\geqslant 1}$ is a sequence of i.i.d. normal random variables, with unknown mean μ_i and unknown variance σ_i^2. The objective is to have a policy π for deciding from which of the N populations to sample from at any time $t = 1,2,\ldots$ so as to maximize the expected sum of outcomes of n total samples or equivalently to minimize the regret due to lack on information of the parameters μ_i and σ_i^2. In this paper, we present a simple inflated sample mean (ISM) index policy that is asymptotically optimal in the sense of Theorem 4 below. This resolves a standing open problem from Burnetas and Katehakis (1996b). Additionally, finite horizon regret bounds are given[1].

Keywords: Inflated Sample Means, Multi-armed Bandits, Sequential Allocation

1. Introduction and Summary

Consider the problem of a controller sampling sequentially from a finite number of $N \geqslant 2$ populations or 'bandits', where the measurements from population i are specified by a sequence of i.i.d. random variables $\{X^i_k\}_{k\geqslant 1}$, taken to be normal with finite mean μ_i and finite variance σ_i^2. The means $\{\mu_i\}$ and variances $\{\sigma_i^2\}$ are taken to be unknown to the controller. It is convenient to define the maximum mean, $\mu^* = \max_i\{\mu_i\}$, and the bandit discrepancies $\{\Delta_i\}$ where $\Delta_i = \mu^* - \mu_i \geqslant 0$. It is additionally convenient to define σ_*^2 as the minimal variance of any bandit that achieves μ^*, that is $\sigma_*^2 = \min_{i:\mu_i=\mu^*} \sigma_i^2$.

In this paper, given k samples from population i we will take the estimators: $\bar{X}^i_k = \sum_{t=1}^k X^i_t/k$ and $S_i^2(k) = \sum_{t=1}^k \left(X^i_t - \bar{X}^i_k\right)^2/k$ for μ_i and σ_i^2 respectively. Note that the use of the biased estimator for the variance, with the $1/k$ factor in place of $1/(k-1)$, is largely for aesthetic purposes - the results presented here adapt to the use of the unbiased estimator as well.

1. Substantial portion of the results reported here were derived independently by Cowan and Katehakis, and by Honda

For any adaptive, non-anticipatory policy π, $\pi(t) = i$ indicates that the controller samples bandit i at time t. Define $T^i_\pi(n) = \sum_{t=1}^n \mathbb{1}\{\pi(t) = i\}$, denoting the number of times bandit i has been sampled during the periods $t = 1, \ldots, n$ under policy π; we take, as a convenience, $T^i_\pi(0) = 0$ for all i, π. The *value* of a policy π is the expected sum of the first n outcomes under π, which we define to be the function $V_\pi(n)$:

$$V_\pi(n) = \mathbb{E}\left[\sum_{i=1}^N \sum_{k=1}^{T^i_\pi(n)} X^i_k\right] = \sum_{i=1}^N \mu_i \mathbb{E}\left[T^i_\pi(n)\right], \tag{1}$$

where for simplicity the dependence of $V_\pi(n)$ on the true, unknown, values of the parameters $\underline{\mu} = (\mu_1, \ldots, \mu_N)$ and $\underline{\sigma}^2 = (\sigma^2_1, \ldots, \sigma^2_N)$, is supressed. The *pseudo-regret*, or simply *regret*, of a policy is taken to be the expected loss due to ignorance of the parameters $\underline{\mu}$ and $\underline{\sigma}^2$ by the controller. Had the controller complete information, she would at every round activate some bandit i^* such that $\mu_{i^*} = \mu^* = \max_i\{\mu_i\}$. For a given policy π, we define the expected regret of that policy at time n as

$$R_\pi(n) = n\mu^* - V_\pi(n) = \sum_{i=1}^n \Delta_i \mathbb{E}\left[T^i_\pi(n)\right]. \tag{2}$$

It follows from Eqs. (1) and (2) that maximization of $V_\pi(n)$with respect to π is equivalent to minimization of $R_\pi(n)$. This type of loss due to ignorance of the means (regret) was first introduced in the context of an $N = 2$ problem by Robbins (1952) as the 'loss per trial' $L_\pi(n)/n = \mu^* - \sum_{i=1}^N \sum_{k=1}^{T^i_\pi(n)} X^i_k/n$ (for which $R_\pi(n) = \mathbb{E}[L_\pi(n)]$), constructing a modified (along two sparse sequences) 'play the winner' policy, π_R, such that $L_{\pi_R}(n) = o(n)$ (a.s.) and $R_{\pi_R}(n) = o(n)$, using for his derivation only the assumption of the Strong Law of Large Numbers. Following Burnetas and Katehakis (1996b) when $n \to \infty$, if π is such that $R_\pi(n) = o(n)$ we say policy π is ***uniformly convergent*** (UC) (since then $\lim_{n\to\infty} V_\pi(n)/n = \mu^*$). However, if under a policy π, $R_\pi(n)$ grew at a slower pace, such as $R_\pi(n) = o(n^{1/2})$, or better $R_\pi(n) = o(n^{1/100})$ etc., then the controller would be assured that π is making a effective trade-off between exploration and exploitation. It turns our that it is possible to construct ***'uniformly fast convergent'*** (UFC) policies, also known as *consistent* or *strongly consistent*, defined as the policies π for which:

$$R_\pi(n) = o(n^\alpha), \text{ for all } \alpha > 0 \text{ for all } (\underline{\mu}, \underline{\sigma}^2).$$

The existence of UFC policies in the case considered here is well established, e.g., Auer et al. (2002) (fig. 4. therein) presented the following UFC policy π_{ACF}:

Policy π_{ACF} (UCB1-NORMAL). At each $n = 1, 2, \ldots$:

i) Sample from any bandit i for which $T^i_{\pi_{\text{ACF}}}(n) < \lceil 8 \ln n \rceil$.

ii) If $T^i_{\pi_{\text{ACF}}}(n) > \lceil 8 \ln n \rceil$, for all $i = 1, \ldots, N$, sample from bandit $\pi_{\text{ACF}}(n+1)$ with

$$\pi_{\text{ACF}}(n+1) = \arg\max_i \left\{ \bar{X}^i_{T^i_\pi(n)} + 4 \cdot S_i(T^i_\pi(n)) \sqrt{\frac{\ln n}{T^i_\pi(n)}} \right\}. \tag{3}$$

(Taking, in this case, $S^2_i(k)$ as the unbiased estimator.)

Additionally, Auer et al. (2002) (in Theorem 4. therein) gave the following bound:

$$R_{\pi_{\text{ACF}}}(n) \leqslant M_{\text{ACF}}(\underline{\mu}, \underline{\sigma}^2) \ln n + C_{\text{ACF}}(\underline{\mu}), \text{ for all } n \text{ and all } (\underline{\mu}, \underline{\sigma}^2), \tag{4}$$

with

$$M_{\mathrm{ACF}}(\underline{\mu},\underline{\sigma}^2) = 256 \sum_{i:\mu_i\neq\mu^*} \frac{\sigma_i^2}{\Delta_i} + 8\sum_{i=1}^{N}\Delta_i, \tag{5}$$

$$C_{\mathrm{ACF}}(\underline{\mu}) = (1+\frac{\pi^2}{2})\sum_{i=1}^{N}\Delta_i. \tag{6}$$

Ineq. (4) readily implies that $R_{\pi_{\mathrm{ACF}}}(n) \leqslant M_{\mathrm{ACF}}(\underline{\mu},\underline{\sigma}^2)\ln n + o(\ln n)$. Thus, since $\ln n = o(n^{\alpha})$ for all $\alpha > 0$ and $R_{\pi_{\mathrm{ACF}}}(n) \geqslant 0$, it follows that π_{ACF} is uniformly fast convergent.

Given that UFC policies exist, the question immediately follows: just how fast can they be? The primary motivation of this paper is the following general result, from Burnetas and Katehakis (1996b), where they showed that for any UFC policy π, the following holds:

$$\liminf_{n\to\infty}\frac{R_\pi(n)}{\ln n} \geqslant \mathbb{M}_{\mathrm{BK}}(\underline{\mu},\underline{\sigma}^2),\ \text{for all } (\underline{\mu},\underline{\sigma}^2), \tag{7}$$

where the bound itself $\mathbb{M}_{\mathrm{BK}}(\underline{\mu},\underline{\sigma}^2)$ is determined by the specific distributions of the populations, in this case

$$\mathbb{M}_{\mathrm{BK}}(\underline{\mu},\underline{\sigma}^2) = \sum_{i:\mu_i\neq\mu^*}\frac{2\Delta_i}{\ln\left(1+\frac{\Delta_i^2}{\sigma_i^2}\right)}\ . \tag{8}$$

For comparison, depending on the specifics of the bandit distributions, there is a considerable distance between the logarithmic term of the upper bound of Eq. (4) and the lower bound implied by Eq. (8).

The derivation of Ineq. (7) implies that in order to guarantee that a policy is uniformly fast convergent, sub-optimal populations have to be sampled at least a logarithmic number of times. The above bound is a special case of a more general result derived in Burnetas and Katehakis (1996b) (part 1 of Theorem 1 therein) for distributions with multi-parameters being unknown (such as in the current problem of Normal populations with both the mean and the variance being unknown):

$$\mathbb{M}_{\mathrm{BK}}(\underline{\mu},\underline{\sigma}^2) = \sum_{i:\mu_i\neq\mu^*}\frac{\Delta_i}{\mathbb{K}_i(\underline{\mu},\underline{\sigma}^2)}$$

with $\mathbb{K}_i(\underline{\mu},\underline{\sigma}^2) = \inf_{(\mu_i',\sigma'^2)}\{\mathbb{I}(f_{(\mu_i,\sigma_i^2)}; f_{(\mu_i',\sigma'^2_i)}) : \mu'_i > \mu^*, \sigma'^2_i > 0\} = (1/2)\ln(1+\frac{\Delta_i^2}{\sigma_i^2})$.

Previously, Lai and Robbins (1985) had obtained such lower bounds for distributions with one-parameter (such as in the current problem of Normal populations with unknown mean but known variance). Allocation policies that achieved the lower bounds were called ***asymptotically efficient*** or ***optimal*** in Lai and Robbins (1985).

Ineq. (7) motivates the definition of a uniformly fast convergent policy π as having a ***uniformly maximal convergence rate*** (UM) or simply being ***asymptotically optimal***, within the class of uniformly fast convergent policies, if $\lim_{n\to\infty} R_\pi(n)/\ln n = \mathbb{M}_{\mathrm{BK}}(\underline{\mu},\underline{\sigma}^2)$, since then $V_\pi(n) = n\mu^* - \mathbb{M}_{\mathrm{BK}}(\underline{\mu},\underline{\sigma}^2)\ln n + o(\ln n)$.

Burnetas and Katehakis (1996b) proposed the following index policy π_{BK} as one that could achieve this lower bound:

Policy π_{BK} (UCB-NORMAL^0)

i) For $n = 1, 2, \ldots, 2N$ sample each bandit twice, and

ii) for $n \geqslant 2N$, sample from bandit $\pi_{\text{BK}}(n+1)$ with

$$\pi_{\text{BK}}(n+1) = \arg\max_i \left\{ \bar{X}^i_{T^i_\pi(n)} + S_i(T^i_\pi(n)) \sqrt{n^{\frac{2}{T^i_\pi(n)}} - 1} \right\}. \tag{9}$$

Burnetas and Katehakis (1996b) were not able to establish the asymptotic optimality of the π_{BK} policy because they were not able to establish a sufficient condition (*Condition* A3 therein), which we express here as the following equivalent conjecture (the referenced open question in the subtitle).

Conjecture 1 *For each i, for every $\varepsilon > 0$, and for $k \to \infty$, the following is true:*

$$\mathbb{P}\left(\bar{X}^i_j + S_i(j)\sqrt{k^{2/j} - 1} < \mu_i - \varepsilon \text{ for some } 2 \leqslant j \leqslant k \right) = o(1/k). \tag{10}$$

We show that the above conjecture is *false* (cf. Proposition 6 in the Appendix). This does not imply that π_{BK} fails to be UM (i.e., to be asymptotically optimal), but this failure means that the techniques established in Burnetas and Katehakis (1996b) are insufficient to verify its optimality. All is not lost, however. One of the central results of this paper is to establish that with a small change, the policy π_{BK} may be modified to one that is provably asymptotically optimal. We introduce in this paper the policy π_{CHK} defined in the following way:

Policy π_{CHK} (UCB-NORMAL^2)

i) For $n = 1, 2, \ldots, 3N$ sample each bandit three times, and

ii) for $n \geqslant 3N$, sample from bandit $\pi_{\text{CHK}}(n+1)$ with

$$\pi_{\text{CHK}}(n+1) = \arg\max_i \left\{ \bar{X}^i_{T^i_\pi(n)} + S_i(T^i_\pi(n)) \sqrt{n^{\frac{2}{T^i_\pi(n)-2}} - 1} \right\}. \tag{11}$$

Remark 1

1) Note that policy π_{CHK} is only a slight modification of policy π_{BK}, the only difference between their indices is the -2 in the power on n under the radical, i.e., $2/(T^i_\pi(n) - 2)$ in $\pi_{\text{CHK}}(n+1)$ replacing $2/T^i_\pi(n)$ in $\pi_{\text{BK}}(n+1)$. This change, while seemingly asymptotically negligible (as in practice $T^i_\pi(n) \to \infty$ (a.s.) with n), has a profound effect on what is provable about π_{CHK}.

2) We note that the indices of policy π_{CHK} are a significant modification of those of the optimal allocation policy $\pi_{\underline{\sigma}^2}$ for the case of normal bandits with *known* variances, cf. Burnetas and Katehakis (1996b) and Katehakis and Robbins (1995), which are:

$$\pi_{\underline{\sigma}^2}(n+1) = \arg\max_i \left\{ \bar{X}^i_{T^i_\pi(n)} + \sigma_i \sqrt{\frac{2\ln n}{T^i_\pi(n)}} \right\}$$

the difference being replacing the term $\sigma_i \sqrt{\frac{2\ln n}{T^i_\pi(n)}}$ in $\pi_{\underline{\sigma}^2}$ by $S_i(T^i_\pi(n))\sqrt{n^{\frac{2}{T^i_\pi(n)-2}} -}$ in π_{CHK}. However, the indices of policy π_{ACF} are a minor modification of the optimal policy π_{σ_i} the difference being replacing the term $\sigma_i \sqrt{\frac{2\ln n}{T^i_\pi(n)}}$ in π_{σ_i} by $S_i(T^i_\pi(n))\sqrt{\frac{16\ln n}{T^i_\pi(n)}}$ in π_{ACF}.

3) The π_{BK} and $\pi_{\underline{\sigma}^2}$ policies can be seen as connected in the following way, however, observing that $2\ln n/T^i_\pi(n)$ is a first-order approximation of $n^{2/T^i_\pi(n)}-1 = e^{2\ln n/T^i_\pi(n)}-1$.

Following Robbins (1952), and additionally Gittins (1979), Lai and Robbins (1985) and Weber (1992) there is a large literature on versions of this problem, cf. Burnetas and Katehakis (2003), Burnetas and Katehakis (1997b) and references therein. For recent work in this area we refer to Audibert et al. (2009), Auer and Ortner (2010), Gittins et al. (2011), Bubeck and Slivkins (2012), Cappé et al. (2013), Kaufmann (2015), Li et al. (2014), Cowan and Katehakis (2015b), Cowan and Katehakis (2015c), and references therein. For more general dynamic programming extensions we refer to Burnetas and Katehakis (1997a), Butenko et al. (2003), Tewari and Bartlett (2008), Audibert et al. (2009), Littman (2012), Feinberg et al. (2014) and references therein. Other related work in this area includes: Burnetas and Katehakis (1993), Burnetas and Katehakis (1996a), Lagoudakis and Parr (2003), Bartlett and Tewari (2009), Tekin and Liu (2012), Jouini et al. (2009), Dayanik et al. (2013), Filippi et al. (2010), Osband and Van Roy (2014), Denardo et al. (2013).

To our knowledge, outside the work in Lai and Robbins (1985), Burnetas and Katehakis (1996b) and Burnetas and Katehakis (1997a), asymptotically optimal policies have only been developed in in Honda and Takemura (2011), and in Honda and Takemura (2010) for the problem of finite known support where optimal policies, cyclic and randomized, that are simpler to implement than those consider in Burnetas and Katehakis (1996b) were constructed. Recently in Cowan and Katehakis (2015a), an asymptotically optimal policy for uniform bandits of unknown support was constructed. The question of whether asymptotically optimal policies exist in the case discussed herein of normal bandits with unknown means and unknown variances was recently resolved in the positive by Honda and Takemura (2013) who demonstrated that a form of Thompson sampling with certain priors on $(\underline{\mu},\underline{\sigma}^2)$ achieves the asymptotic lower bound $\mathbb{M}_{\mathrm{BK}}(\underline{\mu},\underline{\sigma}^2)$.

The structure of the rest of the paper is as follows. In section 2, Theorem 3 establishes a finite horizon bound on the regret of π_{CHK}. From this bound, it follows that π_{CHK} is asymptotically optimal (Theorem 4), and we provide a bound on the remainder term (Theorem 5). Additionally, in Section 3, the Thompson sampling policy of Honda and Takemura (2013) and π_{CHK} are compared and discussed, as both achieve asymptotic optimality.

2. The Optimality Theorem and Finite Time Bounds

The main results of this paper, that Conjecture 1 is false (cf. Proposition 6 in the Appendix), the asymptotic optimality, and the bounds on the behavior of π_{CHK}, all depend on the following probability bounds; we note that tighter bounds seem possible, but these are sufficient for this paper.

Proposition 2 *Let Z, U be independent random variables, $Z\sim N(0,1)$ a standard normal, and $U\sim\chi^2_d$ a chi-squared distribution with d degrees of freedom, where $d\geqslant 2$.*

For $\delta>0,\ p>0$, the following holds for all $k\geqslant 1$:

$$\frac{1}{2}\mathbb{P}\left(\frac{1}{4}Z^2\geqslant U\geqslant\delta^2\right)k^{-d/p}\leqslant\mathbb{P}\left(\delta+\sqrt{U}\sqrt{k^{2/p}-1}<Z\right)\leqslant\frac{e^{-(1+\delta^2)/2}p}{2\delta^2\sqrt{d}}\frac{k^{(1-d)/p}}{\ln k}. \tag{12}$$

Proof [of Proposition 2] The proof is given in the Appendix. ■

Theorem 3 *For policy π_{CHK} as defined above, the following bounds hold for all $n\geqslant 3N$ and all $\varepsilon\in(0,1)$:*

$$R_{\pi_{CHK}}(n)\leqslant\sum_{i:\mu_i\neq\mu^*}\left(\frac{2\ln n}{\ln\left(1+\frac{\Delta_i^2}{\sigma_i^2}\frac{(1-\varepsilon)^2}{(1+\varepsilon)}\right)}+\sqrt{\frac{\pi}{2e}}\frac{8\sigma_*^3}{\Delta_i^3\varepsilon^3}\ln\ln n+\frac{8}{\varepsilon^2}+\frac{8\sigma_i^2}{\Delta_i^2\varepsilon^2}+4\right)\Delta_i. \tag{13}$$

Before giving the proof of this bound, we present two results, the first demonstrating the asymptotic optimality of π_{CHK}, the second giving an ε-free version of the above bound, which gives a bound on the sub-logarithmic remainder term. It is worth noting the following. The bounds of Theorem 3 can actually be improved, through the use of a modified version of Proposition 2, to eliminate the $\ln\ln n$ dependence, so the only dependence on n is through the initial $\ln n$ term. The cost of this, however, is a dependence on a larger power of $1/\varepsilon$. The particular form of the bound given in Eq. (13) was chosen to simplify the following two results, cf. Remark 4 in the proof of Propositition 2.

Theorem 4 *For a policy π_{CHK} as defined above, π_{CHK} is asymptotically optimal in the sense that*

$$\lim_{n\to\infty} \frac{R_{\pi_{CHK}}(n)}{\ln n} = \mathbb{M}_{BK}(\underline{\mu}, \underline{\sigma}^2). \tag{14}$$

Proof [of Theorem 4] For any ε such that $0 < \varepsilon < 1$, we have from Theorem 3 that the followings holds:

$$\limsup_{n\to\infty} \frac{R_{\pi_{\text{CHK}}}(n)}{\ln n} \leqslant \sum_{i:\mu_i \neq \mu^*} \frac{2\Delta_i}{\ln\left(1 + \frac{\Delta_i^2}{\sigma_i^2}\frac{(1-\varepsilon)^2}{(1+\varepsilon)}\right)}. \tag{15}$$

Taking the infimum over all such ε,

$$\limsup_{n\to\infty} \frac{R_{\pi_{\text{CHK}}}(n)}{\ln n} \leqslant \sum_{i:\mu_i \neq \mu^*} \frac{2\Delta_i}{\ln\left(1 + \frac{\Delta_i^2}{\sigma_i^2}\right)} = \mathbb{M}_{\text{BK}}(\underline{\mu}, \underline{\sigma}^2), \tag{16}$$

and observing the lower bound of Eq. (7) completes the result. ■

Theorem 5 *For a policy π_{CHK} as defined above, $R_{\pi_{CHK}}(n) \leqslant \mathbb{M}_{BK}(\underline{\mu}, \underline{\sigma}^2)\ln n + O((\ln n)^{3/4}\ln\ln n)$, and more concretely*

$$\begin{aligned} R_{\pi_{CHK}}(n) \leqslant M^0_{CHK}(\underline{\mu}, \underline{\sigma}^2)\ln n &+ M^1_{CHK}(\underline{\mu}, \underline{\sigma}^2)(\ln n)^{3/4}\ln\ln n \\ &+ M^2_{CHK}(\underline{\mu}, \underline{\sigma}^2)(\ln n)^{3/4} \\ &+ M^3_{CHK}(\underline{\mu}, \underline{\sigma}^2)(\ln n)^{1/2} \\ &+ M^4_{CHK}(\underline{\mu}, \underline{\sigma}^2), \end{aligned} \tag{17}$$

where

$$\begin{aligned} M^0_{CHK}(\underline{\mu}, \underline{\sigma}^2) &= \mathbb{M}_{BK}(\underline{\mu}, \underline{\sigma}^2) \\ M^1_{CHK}(\underline{\mu}, \underline{\sigma}^2) &= 64\sqrt{\frac{\pi}{2e}} \sum_{i:\mu_i \neq \mu^*} \left(\frac{\sigma_*^3}{\Delta_i^2}\right) \\ M^2_{CHK}(\underline{\mu}, \underline{\sigma}^2) &= 10 \sum_{i:\mu_i \neq \mu^*} \left(\frac{\Delta_i^3}{(\sigma_i^2 + \Delta_i^2)\ln\left(1 + \frac{\Delta_i^2}{\sigma_i^2}\right)^2}\right) \\ M^3_{CHK}(\underline{\mu}, \underline{\sigma}^2) &= 32 \sum_{i:\mu_i \neq \mu^*} \left(\Delta_i + \frac{\sigma_i^2}{\Delta_i}\right) \\ M^4_{CHK}(\underline{\mu}, \underline{\sigma}^2) &= 4 \sum_{i:\mu_i \neq \mu^*} \Delta_i. \end{aligned} \tag{18}$$

While the above bound admittedly has a more complex form than such a bound as in Eq. (4), it demonstrates the asymptotic optimality of the dominating term, and bounds the sub-linear remainder term.

Proof [of Theorem 5] The bound follows directly from Theorem 3, taking $\varepsilon = \frac{1}{2}(\ln n)^{-1/4}$ for $n \geqslant 3$, and observing the following bound, that for ε such that $0 < \varepsilon < 1/2$,

$$\frac{1}{\ln\left(1+\frac{\Delta_i^2}{\sigma_i^2}\frac{(1-\varepsilon)^2}{1+\varepsilon}\right)} \leqslant \frac{1}{\ln\left(1+\frac{\Delta_i^2}{\sigma_i^2}\right)} + \frac{10\Delta_i^2}{(\sigma_i^2+\Delta_i^2)\ln\left(1+\frac{\Delta_i^2}{\sigma_i^2}\right)^2}\varepsilon. \tag{19}$$

This inequality is proven separately as Proposition 7 in the Appendix. ■

We make no claim that the results of Theorems 3, 5 are the best achievable for this policy π_{CHK}. At several points in the proofs, choices of convenience were made in the bounding of terms, and different techniques may yield tighter bounds still. But they are sufficient to demonstrate the asymptotic optimality of π_{CHK}, and give useful bounds on the growth of $R_{\pi_{\text{CHK}}}(n)$.

Proof [of Theorem 1] In this proof, we take $\pi = \pi_{\text{CHK}}$ as defined above. For notational convenience, we define the index function

$$u_i(k,j) = \bar{X}_j^i + S_i(j)\sqrt{k^{\frac{2}{j-2}}-1}. \tag{20}$$

The structure of this proof will be to bound the expected value of $T_\pi^i(n)$ for all sub-optimal bandits i, and use this to bound the regret $R_\pi(n)$. The basic techniques follow those in Katehakis and Robbins (1995) for the known variance case, modified accordingly here for the unknown variance case and assisted by the probability bound of Proposition 2. For any i such that $\mu_i \neq \mu^*$, we define the following quantities: Let $1 > \varepsilon > 0$ and define $\tilde{\varepsilon} = \Delta_i\varepsilon/2$. For $n \geqslant 3N$,

$$\begin{aligned}
n_1^i(n,\varepsilon) &= \sum_{t=3N}^{n} \mathbb{1}\{\pi(t+1)=i, u_i(t,T_\pi^i(t)) \geqslant \mu^*-\tilde{\varepsilon}, \bar{X}^i_{T_\pi^i(t)} \leqslant \mu_i+\tilde{\varepsilon}, S_i^2(T_\pi^i(t)) \leqslant \sigma_i^2(1+\varepsilon)\} \\
n_2^i(n,\varepsilon) &= \sum_{t=3N}^{n} \mathbb{1}\{\pi(t+1)=i, u_i(t,T_\pi^i(t)) \geqslant \mu^*-\tilde{\varepsilon}, \bar{X}^i_{T_\pi^i(t)} \leqslant \mu_i+\tilde{\varepsilon}, S_i^2(T_\pi^i(t)) > \sigma_i^2(1+\varepsilon)\} \\
n_3^i(n,\varepsilon) &= \sum_{t=3N}^{n} \mathbb{1}\{\pi(t+1)=i, u_i(t,T_\pi^i(t)) \geqslant \mu^*-\tilde{\varepsilon}, \bar{X}^i_{T_\pi^i(t)} > \mu_i+\tilde{\varepsilon}\} \\
n_4^i(n,\varepsilon) &= \sum_{t=3N}^{n} \mathbb{1}\{\pi(t+1)=i, u_i(t,T_\pi^i(t)) < \mu^*-\tilde{\varepsilon}\}.
\end{aligned} \tag{21}$$

Hence, we have the following relationship for $n \geqslant 3N$, that

$$T_\pi^i(n+1) = 3 + \sum_{t=3N}^{n} \mathbb{1}\{\pi(t+1)=i\} = 3 + n_1^i(n,\varepsilon) + n_2^i(n,\varepsilon) + n_3^i(n,\varepsilon) + n_4^i(n,\varepsilon). \tag{22}$$

The proof proceeds by bounding, in expectation, each of the four terms.

Observe that, by the structure of the index function u_i,

$$
\begin{aligned}
n_1^i(n,\varepsilon) &\leqslant \sum_{t=3N}^{n} \mathbb{1}\left\{\pi(t+1)=i, (\mu_i+\tilde{\varepsilon})+\sigma_i\sqrt{1+\varepsilon}\sqrt{t^{\frac{2}{T_\pi^i(t)-2}}-1} \geqslant \mu^*-\tilde{\varepsilon}\right\} \\
&= \sum_{t=3N}^{n} \mathbb{1}\left\{\pi(t+1)=i, T_\pi^i(t) \leqslant \frac{2\ln t}{\ln\left(1+\frac{1}{\sigma_i^2}\frac{(\Delta_i-2\tilde{\varepsilon})^2}{(1+\varepsilon)}\right)}+2\right\} \\
&= \sum_{t=3N}^{n} \mathbb{1}\left\{\pi(t+1)=i, T_\pi^i(t) \leqslant \frac{2\ln t}{\ln\left(1+\frac{\Delta_i^2}{\sigma_i^2}\frac{(1-\varepsilon)^2}{(1+\varepsilon)}\right)}+2\right\} \\
&\leqslant \sum_{t=3N}^{n} \mathbb{1}\left\{\pi(t+1)=i, T_\pi^i(t) \leqslant \frac{2\ln n}{\ln\left(1+\frac{\Delta_i^2}{\sigma_i^2}\frac{(1-\varepsilon)^2}{(1+\varepsilon)}\right)}+2\right\} \\
&\leqslant \sum_{t=1}^{n} \mathbb{1}\left\{\pi(t+1)=i, T_\pi^i(t) \leqslant \frac{2\ln n}{\ln\left(1+\frac{\Delta_i^2}{\sigma_i^2}\frac{(1-\varepsilon)^2}{(1+\varepsilon)}\right)}+2\right\} \\
&\leqslant \frac{2\ln n}{\ln\left(1+\frac{\Delta_i^2}{\sigma_i^2}\frac{(1-\varepsilon)^2}{(1+\varepsilon)}\right)}+2+2.
\end{aligned} \tag{23}
$$

The last inequality follows, observing that $T_\pi^i(t)$ may be expressed as the sum of $\pi(t)=i$ indicators, and seeing that the additional condition bounds the number of non-zero terms in the above sum. The additional $+2$ simply accounts for the $\pi(1)=i$ term and the $\pi(n+1)=i$ term. Note, this bound is sample-path-wise.

For the second term,

$$
\begin{aligned}
n_2^i(n,\varepsilon) &\leqslant \sum_{t=3N}^{n} \mathbb{1}\{\pi(t+1)=i, S_i^2(T_\pi^i(t)) > \sigma_i^2(1+\varepsilon)\} \\
&= \sum_{t=3N}^{n}\sum_{k=2}^{t} \mathbb{1}\{\pi(t+1)=i, S_i^2(k) > \sigma_i^2(1+\varepsilon), T_\pi^i(t)=k\} \\
&= \sum_{t=3N}^{n}\sum_{k=2}^{t} \mathbb{1}\{\pi(t+1)=i, T_\pi^i(t)=k\}\mathbb{1}\{S_i^2(k) > \sigma_i^2(1+\varepsilon)\} \\
&\leqslant \sum_{k=2}^{n} \mathbb{1}\{S_i^2(k) > \sigma_i^2(1+\varepsilon)\}\sum_{t=k}^{n}\mathbb{1}\{\pi(t+1)=i, T_\pi^i(t)=k\} \\
&\leqslant \sum_{k=2}^{n} \mathbb{1}\{S_i^2(k) > \sigma_i^2(1+\varepsilon)\}.
\end{aligned} \tag{24}
$$

The last inequality follows as, for fixed k, $\{\pi(t+1)=i, T_\pi^i(t)=k\}$ may be true for at most one value of t. Recall that $kS_i^2(k)/\sigma_i^2$ has the distribution of a χ^2_{k-1} random variable. Letting $U_k \sim \chi^2_k$, from the above we

have

$$
\begin{aligned}
\mathbb{E}\left[n_2^i(n,\varepsilon)\right] &\leqslant \sum_{k=2}^{n} \mathbb{P}\left(S_i^2(k) > \sigma_i^2(1+\varepsilon)\right) \\
&\leqslant \sum_{k=2}^{\infty} \mathbb{P}\left(U_{k-1}/k > (1+\varepsilon)\right) \\
&\leqslant \sum_{k=2}^{\infty} \mathbb{P}\left(U_{k-1}/(k-1) > (1+\varepsilon)\right) \\
&= \sum_{k=1}^{\infty} \mathbb{P}\left(U_k > k(1+\varepsilon)\right) \\
&\leqslant \frac{1}{\sqrt{\frac{e^{\varepsilon}}{1+\varepsilon}}-1} \leqslant \frac{8}{\varepsilon^2} < \infty.
\end{aligned} \tag{25}
$$

The penultimate step is a Chernoff bound on the terms, $\mathbb{P}\left(U_k > k(1+\varepsilon)\right) \leqslant (e^{-\varepsilon}(1+\varepsilon))^{k/2}$.

To bound the third term, a similar rearrangement to Eq. (24) (using the sample mean instead of the sample variance) yields:

$$
n_3^i(n,\varepsilon) \leqslant \sum_{t=3N}^{n} \mathbb{1}\{\pi(t+1)=i, \bar{X}^i_{T^i_\pi(t)} > \mu_i + \tilde{\varepsilon}\} \leqslant \sum_{k=2}^{n} \mathbb{1}\{\bar{X}^i_k > \mu_i + \tilde{\varepsilon}\}. \tag{26}
$$

Recalling that $\bar{X}^i_k - \mu_i \sim Z\sigma_i/\sqrt{k}$ for Z a standard normal,

$$
\mathbb{E}\left[n_3^i(n,\varepsilon)\right] \leqslant \sum_{k=2}^{n} \mathbb{P}\left(\bar{X}^i_k > \mu_i + \tilde{\varepsilon}\right) \leqslant \sum_{k=1}^{\infty} \mathbb{P}\left(Z\sigma_i/\sqrt{k} > \tilde{\varepsilon}\right) \leqslant \frac{1}{e^{\frac{\tilde{\varepsilon}^2}{2\sigma_i^2}}-1} \leqslant \frac{2\sigma_i^2}{\tilde{\varepsilon}^2} < \infty. \tag{27}
$$

The penultimate step is a Chernoff bound on the terms, $\mathbb{P}\left(Z > \delta\sqrt{k}\right) \leqslant e^{-k\delta^2/2}$.

To bound the n_4^i term, observe that in the event $\pi(t+1)=i$, from the structure of the policy it must be true that $u_i(t, T^i_\pi(t)) = \max_j u_j(t, T^j_\pi(t))$. Thus, if i^* is some bandit such that $\mu_{i^*} = \mu^*$, $u_{i^*}(t, T^{i^*}_\pi(t)) \leqslant u_i(t, T^i_\pi(t))$. In particular, we take i^* to be a bandit that not only achieves the maximal mean μ^*, but also the minimal variance among optimal bandits, $\sigma^2_{i^*} = \sigma^2_*$. We have the following bound,

$$
\begin{aligned}
n_4^i(n,\varepsilon) &\leqslant \sum_{t=3N}^{n} \mathbb{1}\{\pi(t+1)=i, u_{i^*}(t, T^{i^*}_\pi(t)) < \mu^* - \tilde{\varepsilon}\} \\
&\leqslant \sum_{t=3N}^{n} \mathbb{1}\{u_{i^*}(t, T^{i^*}_\pi(t)) < \mu^* - \tilde{\varepsilon}\} \\
&\leqslant \sum_{t=3N}^{n} \mathbb{1}\{u_{i^*}(t,s) < \mu^* - \tilde{\varepsilon} \text{ for some } 3 \leqslant s \leqslant t\}.
\end{aligned} \tag{28}
$$

The last step follows as for t in this range, $3 \leqslant T^{i^*}_\pi(t) \leqslant t$. Hence

$$
\mathbb{E}\left[n_4^i(n,\varepsilon)\right] \leqslant \sum_{t=3N}^{n} \mathbb{P}\left(u_{i^*}(t,s) < \mu^* - \tilde{\varepsilon} \text{ for some } 3 \leqslant s \leqslant t\right). \tag{29}
$$

As an aside, this is essentially the point at which the conjectured Eq. (10) would have come into play for the proof of the optimality of π_{BK}, bounding the growth of the corresponding term for that policy. We will essentially prove a successful version of that conjecture here. Define the events $A^*_{s,t,\varepsilon} = \{u_{i^*}(t,s) < \mu^* - \tilde{\varepsilon}\}$.

Observing the distributions of the sample mean and sample variance, we have (similar to Eq. (41)) for Z a standard normal and $U_{s-1} \sim \chi^2_{s-1}$, with U, Z independent,

$$
\begin{aligned}
\mathbb{P}\left(A^*_{s,t,\varepsilon}\right) &= \mathbb{P}\left(\frac{\tilde{\varepsilon}}{\sigma_*}\sqrt{s} + \sqrt{U_{s-1}}\sqrt{t^{\frac{2}{s-2}} - 1} < Z\right) \\
&\leqslant \frac{e^{-(\tilde{\varepsilon}/\sigma_*)^2 s/2}(s-2)}{2(\tilde{\varepsilon}/\sigma_*)^2 s\sqrt{e(s-1)}}\left(\frac{t^{-1}}{\ln t}\right) \\
&\leqslant \frac{e^{-(\tilde{\varepsilon}/\sigma_*)^2 s/2}}{2(\tilde{\varepsilon}/\sigma_*)^2}\frac{1}{\sqrt{es}}\left(\frac{t^{-1}}{\ln t}\right) \\
&\leqslant \left(\frac{1}{2(\tilde{\varepsilon}/\sigma_*)^2\sqrt{e}}\right)\frac{e^{-(\tilde{\varepsilon}/\sigma_*)^2 s/2}}{\sqrt{s}}\left(\frac{t^{-1}}{\ln t}\right).
\end{aligned}
\tag{30}
$$

where the first inequality follows as an application of Proposition 2, and the second since $s \geqslant 3$. Applying a union bound to Eq. (29),

$$
\begin{aligned}
\mathbb{E}\left[n^i_4(n,\varepsilon)\right] &\leqslant \sum_{t=3N}^{n}\sum_{s=3}^{t}\mathbb{P}\left(A^*_{s,t,\varepsilon}\right) \\
&\leqslant \sum_{t=3N}^{n}\sum_{s=3}^{t}\left(\frac{1}{2(\tilde{\varepsilon}/\sigma_*)^2\sqrt{e}}\right)\frac{e^{-(\tilde{\varepsilon}/\sigma_*)^2 s/2}}{\sqrt{s}}\left(\frac{t^{-1}}{\ln t}\right) \\
&\leqslant \left(\frac{1}{2(\tilde{\varepsilon}/\sigma_*)^2\sqrt{e}}\right)\int_{s=0}^{\infty}\frac{e^{-(\tilde{\varepsilon}/\sigma_*)^2 s/2}}{\sqrt{s}}ds\int_{t=e}^{n}\left(\frac{t^{-1}}{\ln t}\right)dt \\
&= \left(\frac{1}{2(\tilde{\varepsilon}/\sigma_*)^2\sqrt{e}}\right)\frac{\sqrt{2\pi}}{(\tilde{\varepsilon}/\sigma_*)}\ln\ln n \\
&= \sqrt{\frac{\pi}{2e}}\frac{\sigma_*^3}{\tilde{\varepsilon}^3}\ln\ln n.
\end{aligned}
\tag{31}
$$

The bounds follow, removing the dependence of the s-sum on t by extending it to ∞, and bounding the sums by integrals of the (decreasing) summands by slightly extending the range of each. From the above results, and observing that $T^i_\pi(n) \leqslant T^i_\pi(n+1)$, it follows from Eq. (22) that for any ε such that $0 < \varepsilon < 1$,

$$
\begin{aligned}
\mathbb{E}\left[T^i_\pi(n)\right] &\leqslant \frac{2\ln n}{\ln\left(1+\frac{\Delta_i^2}{\sigma_i^2}\frac{(1-\varepsilon)^2}{(1+\varepsilon)}\right)} + 4 + \frac{8}{\varepsilon^2} + \frac{2\sigma_i^2}{\tilde{\varepsilon}^2} + \sqrt{\frac{\pi}{2e}}\frac{\sigma_i^3}{\tilde{\varepsilon}^3}\ln\ln n \\
&\leqslant \frac{2\ln n}{\ln\left(1+\frac{\Delta_i^2}{\sigma_i^2}\frac{(1-\varepsilon)^2}{(1+\varepsilon)}\right)} + 4 + \frac{8}{\varepsilon^2} + \frac{8\sigma_i^2}{\Delta_i^2\varepsilon^2} + \sqrt{\frac{\pi}{2e}}\frac{8\sigma_*^3}{\Delta_i^3\varepsilon^3}\ln\ln n.
\end{aligned}
\tag{32}
$$

The result then follows from the definition of regret in Eq. (2).

■

Remark 2 Numerical Regret Comparison: Figure 1 shows the results of a small simulation study done on a set of six populations with means and variances given in Table 1. It provides plots of the regrets when implementing policies $\pi_{\text{CHK}}, \pi_{\text{ACF}}$, and π_G a 'greedy' policy that always activates the bandit with the current highest average. Each policy was implemented over a horizon of 100,000 activations, each replicated 10,000 times to produce a good estimate of the average regret $R_\pi(n)$ over the times indicated. The left plot is on the time scale of the first $10,000$ activations, and the right is on the full time scale of $100,000$ activations.

μ_i	8	8	7.9	7	-1	0
σ_i^2	1	1.4	0.5	3	1	4

Table 1

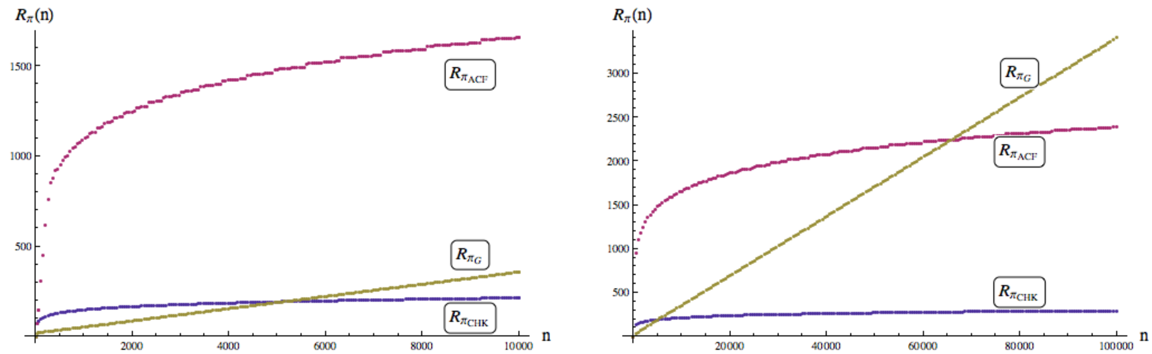

Figure 1: Numerical Regret Comparison of $\pi_{\rm ACF}$, $\pi_{\rm CHK}$, and $\pi_{\rm G}$; Left: $[0, 10,000]$ range, Right: $[0, 100,000]$ range.

Remark 3 Bounds and Limits: Figure 2 shows first (left) a comparison of the theoretical bounds on the regret, $B_{\pi_{\rm ACF}}(n)$ and $B_{\pi_{\rm CHK}}(n)$ representing the theoretical regret bounds of the RHS of Eq. (4) and Eq. (13) respectively, taking $\varepsilon = (\ln n)^{1/4}$ in the latter case, for the means and variances indicated in Table 1. Additionally, Figure 2 (right) shows the convergence of $R_{\pi_{\rm CHK}}(n)/\ln n$ to the theoretical lower bound $\mathbb{M}_{\rm BK}(\underline{\mu}, \underline{\sigma}^2)$.

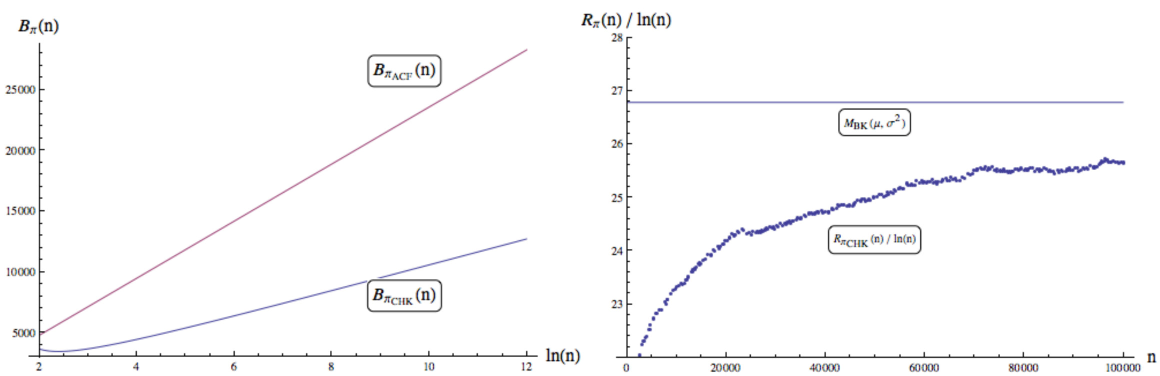

Figure 2: Left: Plots of $B_{\pi_{\rm ACF}}(n)$ and $B_{\pi_{\rm CHK}}(n)$. Right: Convergence of $R_{\pi_{\rm CHK}}(n)/\ln(n)$ to $\mathbb{M}_{\rm BK}(\underline{\mu}, \underline{\sigma}^2)$

3. A Comparison of $\pi_{\rm CHK}$and Thompson Sampling

Honda and Takemura (2013) proved that for $\alpha < 0$, the following Thompson sampling algorithm is asymptotically optimal, i.e., $\lim_{n\to\infty} R_{\pi_{\rm CHK}}(n)/\ln n = \mathbb{M}_{\rm BK}(\underline{\mu}, \underline{\sigma}^2)$.

Policy π_{TS} (TS-NORMAL$^{\alpha}$)

i) Initially, sample each bandit $\tilde{n} \geqslant \max(2, 3 - \lfloor 2\alpha \rfloor)$ times.

ii) For $n \geqslant \tilde{n}$: For each i generate a random sample U_n^i from a posterior distribution for μ_i, given $\left(\bar{X}^i_{T^i_\pi(n)}, S_i^2(T^i_\pi(n))\right)$, and a prior for $\left(\mu_i, \sigma_i^2\right) \propto \left(\sigma_i^2\right)^{-1-\alpha}$.

iii) Then, take

$$\pi_{\text{TS}}(n+1) = \arg\max_i U_n^i. \tag{33}$$

Policies π_{TS} and π_{CHK} differ decidedly in structure. One key difference, π_{TS} is an inherently randomized policy, while decisions under π_{CHK} are completely determined given the bandit results at a given time. Given that both π_{TS} and π_{CHK} are asymptotically optimal, it is interesting to compare the performances of these two algorithms over finite time horizons, and observe any practical differences between them. To that end, two small simulation studies were done for different sets of bandit parameters $(\underline{\mu}, \underline{\sigma}^2)$. In each case, the uniform prior $\alpha = -1$ was used. The simulations were carried out on a 10,000 round time horizon, and replicated sufficiently many times to get good estimates for the expected regret over the times indicated.

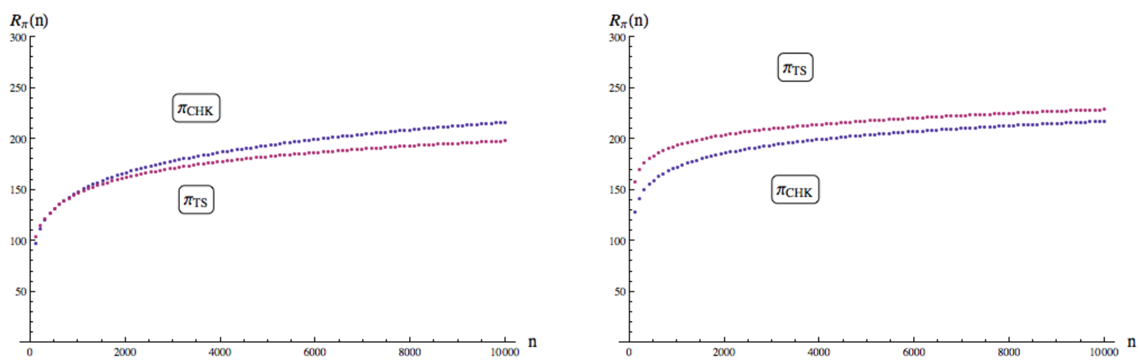

Figure 3: Numerical Regret Comparison of $\pi_{\text{CHK}}$and π_{TS} for the parameters, of Table 1, left and Table 2, right.

μ_i	10	9	8	7	-1	0
σ_i^2	8	1	1	0.5	1	4

Table 2

We observe from the above, and from general sampling of bandit parameters, that π_{TS} and π_{CHK} generally produce comparable expected regret. A general exploration of random parameters suggests that, on average, π_{TS} is slightly superior to π_{CHK} in cases where all bandits have roughly equal variances, while π_{CHK} has an edge when the optimal bandits have large variance relative to the other bandits, and the size of the bandit discrepancies. It is additionally interesting to note that in the cases pictured above, the superior policy also demonstrated the smaller variance in sample regret (Figure 4). Additional numerical experiments, not pictured here, indicate that the superior policy in each case may exhibit a slightly heavier tail distribution towards larger regret. In general, the question of which policy is superior is largely context specific.

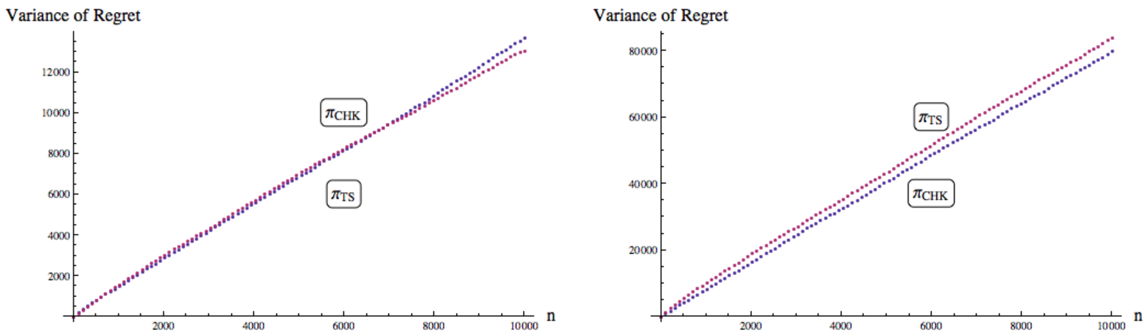

Figure 4: Numerical comparison of variance of sample regret for π_{CHK} and π_{TS} for different parameters, of Table 1, left and Table 2, right.

References

Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. Exploration–exploitation tradeoff using variance estimates in multi-armed bandits. *Theoretical Computer Science*, 410(19):1876–1902, 2009.

Peter Auer and Ronald Ortner. Ucb revisited: Improved regret bounds for the stochastic multi-armed bandit problem. *Periodica Mathematica Hungarica*, 61(1-2):55–65, 2010.

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Peter L Bartlett and Ambuj Tewari. Regal: A regularization based algorithm for reinforcement learning in weakly communicating mdps. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 35–42. AUAI Press, 2009.

Sébastien Bubeck and Aleksandrs Slivkins. The best of both worlds: Stochastic and adversarial bandits. arXiv preprint arXiv:1202.4473, 2012.

Apostolos N Burnetas and Michael N Katehakis. On sequencing two types of tasks on a single processor under incomplete information. *Probability in the Engineering and Informational Sciences*, 7(1):85–119, 1993.

Apostolos N Burnetas and Michael N Katehakis. On large deviations properties of sequential allocation problems. *Stochastic Analysis and Applications*, 14(1):23–31, 1996a.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for sequential allocation problems. *Advances in Applied Mathematics*, 17(2):122–142, 1996b.

Apostolos N Burnetas and Michael N Katehakis. Optimal adaptive policies for Markov decision processes. *Mathematics of Operations Research*, 22(1):222–255, 1997a.

Apostolos N Burnetas and Michael N Katehakis. On the finite horizon one-armed bandit problem. *Stochastic Analysis and Applications*, 16(1):845–859, 1997b.

Apostolos N Burnetas and Michael N Katehakis. Asymptotic Bayes analysis for the finite-horizon one-armed-bandit problem. *Probability in the Engineering and Informational Sciences*, 17(01):53–82, 2003.

Sergiy Butenko, Panos M Pardalos, and Robert Murphey. *Cooperative Control: Models, Applications, and Algorithms*. Kluwer Academic Publishers, 2003.

Olivier Cappé, Aurélien Garivier, Odalric-Ambrym Maillard, Rémi Munos, and Gilles Stoltz. Kullback–leibler upper confidence bounds for optimal sequential allocation. *The Annals of Statistics*, 41(3):1516–1541, 2013.

Wesley Cowan and Michael N Katehakis. An asymptotically optimal UCB policy for uniform bandits of unknown support. arXiv preprint arXiv:1505.01918, 2015a.

Wesley Cowan and Michael N Katehakis. Asymptotic behavior of minimal-exploration allocation policies: Almost sure, arbitrarily slow growing regret. arXiv preprint arXiv:1505.02865, Jul. 31 2015b.

Wesley Cowan and Michael N Katehakis. Multi-armed bandits under general depreciation and commitment. *Probability in the Engineering and Informational Sciences*, 29(01):51–76, 2015c.

Savas Dayanik, Warren B Powell, and Kazutoshi Yamazaki. Asymptotically optimal Bayesian sequential change detection and identification rules. *Annals of Operations Research*, 208(1):337–370, 2013.

Eric V Denardo, Eugene A Feinberg, and Uriel G Rothblum. The multi-armed bandit, with constraints. In M.N. Katehakis, S.M. Ross, and J. Yang, editors, *Cyrus Derman Memorial Volume I: Optimization under Uncertainty: Costs, Risks and Revenues*. Annals of Operations Research, Springer, New York, 2013.

Eugene A Feinberg, Pavlo O Kasyanov, and Michael Z Zgurovsky. Convergence of value iterations for total-cost mdps and pomdps with general state and action sets. In *Adaptive Dynamic Programming and Reinforcement Learning (ADPRL), 2014 IEEE Symposium on*, pages 1–8. IEEE, 2014.

Sarah Filippi, Olivier Cappé, and Aurélien Garivier. Optimism in reinforcement learning based on kull-backleibler divergence. In *48th Annual Allerton Conference on Communication, Control, and Computing*, 2010.

John C. Gittins. Bandit processes and dynamic allocation indices (with discussion). *J. Roy. Stat. Soc. Ser. B*, 41:335–340, 1979.

John C. Gittins, Kevin Glazebrook, and Richard R. Weber. *Multi-armed Bandit Allocation Indices*. John Wiley & Sons, West Sussex, U.K., 2011.

Junya Honda and Akimichi Takemura. An asymptotically optimal bandit algorithm for bounded support models. In *COLT*, pages 67–79. Citeseer, 2010.

Junya Honda and Akimichi Takemura. An asymptotically optimal policy for finite support models in the multiarmed bandit problem. *Machine Learning*, 85(3):361–391, 2011.

Junya Honda and Akimichi Takemura. Optimality of Thompson sampling for Gaussian bandits depends on priors. arXiv preprint arXiv:1311.1894, 2013.

Wassim Jouini, Damien Ernst, Christophe Moy, and Jacques Palicot. Multi-armed bandit based policies for cognitive radio's decision making issues. In *3rd international conference on Signals, Circuits and Systems (SCS)*, 2009.

Michael N Katehakis and Herbert Robbins. Sequential choice from several populations. *Proceedings of the National Academy of Sciences of the United States of America*, 92(19):8584, 1995.

Emilie Kaufmann. Analyse de stratégies Bayésiennes et fréquentistes pour l'allocation séquentielle de ressources. *Doctorat*, ParisTech., Jul. 31 2015.

Michail G Lagoudakis and Ronald Parr. Least-squares policy iteration. *The Journal of Machine Learning Research*, 4:1107–1149, 2003.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985.

Lihong Li, Remi Munos, and Csaba Szepesvari. On minimax optimal offline policy evaluation. arXiv preprint arXiv:1409.3653, 2014.

Michael L Littman. Inducing partially observable Markov decision processes. In *ICGI*, pages 145–148, 2012.

Ian Osband and Benjamin Van Roy. Near-optimal reinforcement learning in factored mdps. In *Advances in Neural Information Processing Systems*, pages 604–612, 2014.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bull. Amer. Math. Monthly*, 58: 527–536, 1952.

Cem Tekin and Mingyan Liu. Approximately optimal adaptive learning in opportunistic spectrum access. In *INFOCOM, 2012 Proceedings IEEE*, pages 1548–1556. IEEE, 2012.

Ambuj Tewari and Peter L Bartlett. Optimistic linear programming gives logarithmic regret for irreducible mdps. In *Advances in Neural Information Processing Systems*, pages 1505–1512, 2008.

Richard R Weber. On the Gittins index for multiarmed bandits. *The Annals of Applied Probability*, 2(4): 1024–1033, 1992.

Acknowledgement: We gratefully acknowledge support for this project from the National Science Foundation (NSF grant CMMI-14-50743).

Appendix A. Additional Proofs

Proof [of Proposition 2] Let $P = \mathbb{P}\left(\delta + \sqrt{U}\sqrt{k^{2/p}-1} < Z\right)$. Note immediately, $P \geqslant \mathbb{P}\left(\delta + \sqrt{U}k^{1/p} < Z\right)$. Further,

$$
\begin{aligned}
P &\geqslant \mathbb{P}\left(\delta + \sqrt{U}k^{1/p} < Z \text{ and } \sqrt{U}k^{1/p} \geqslant \delta\right) \\
&\geqslant \mathbb{P}\left(2\sqrt{U}k^{1/p} < Z \text{ and } \sqrt{U}k^{1/p} \geqslant \delta\right) \\
&= \int_{\frac{\delta^2}{k^{2/p}}}^{\infty} \int_{2\sqrt{u}k^{1/p}}^{\infty} \frac{e^{-z^2/2}}{\sqrt{2\pi}} f_d(u)\,dz\,du.
\end{aligned}
\tag{34}
$$

Where $f_d(u)$ is taken to be the density of a χ^2_d-random variable. Letting $\tilde{u} = k^{2/p}u$,

$$
\begin{aligned}
P &\geqslant \frac{1}{k^{2/p}} \int_{\delta^2}^{\infty} \int_{2\sqrt{\tilde{u}}}^{\infty} \frac{e^{-z^2/2}}{\sqrt{2\pi}} f_d\left(\frac{\tilde{u}}{k^{2/p}}\right) dz\,d\tilde{u} \\
&= \frac{1}{k^{2/p}} \int_{\delta^2}^{\infty} \int_{2\sqrt{\tilde{u}}}^{\infty} \frac{e^{-z^2/2}}{\sqrt{2\pi}} \frac{1}{2^{d/2}\Gamma(d/2)} \left(\frac{\tilde{u}}{k^{2/p}}\right)^{d/2-1} e^{-\frac{\tilde{u}}{2k^{2/p}}}\,dz\,d\tilde{u} \\
&= \left(\frac{1}{k^{2/p}}\right)^{d/2} \int_{\delta^2}^{\infty} \int_{2\sqrt{\tilde{u}}}^{\infty} \frac{e^{-z^2/2}}{\sqrt{2\pi}} \frac{1}{2^{d/2}\Gamma(d/2)} \tilde{u}^{d/2-1} e^{-\frac{\tilde{u}}{2k^{2/p}}}\,dz\,d\tilde{u}
\end{aligned}
\tag{35}
$$

Observing that $k^{2/p} \geqslant 1$,

$$\begin{aligned} P &\geqslant \left(\frac{1}{k^{2/p}}\right)^{d/2} \int_{\delta^2}^{\infty} \int_{2\sqrt{\tilde{u}}}^{\infty} \frac{e^{-z^2/2}}{\sqrt{2\pi}} \frac{1}{2^{d/2}\Gamma(d/2)} \tilde{u}^{d/2-1} e^{-\frac{\tilde{u}}{2}} dz d\tilde{u} \\ &= k^{-d/p} \mathbb{P}\left(2\sqrt{U} \leqslant Z \text{ and } U \geqslant \delta^2\right) \\ &= \frac{1}{2} k^{-d/p} \mathbb{P}\left(4U \leqslant Z^2 \text{ and } U \geqslant \delta^2\right) = \frac{1}{2} k^{-d/p} \mathbb{P}\left(\frac{1}{4} Z^2 \geqslant U \geqslant \delta^2\right) \end{aligned} \tag{36}$$

The exchange from integral to probability is simply the interpretation of the integrand as the joint pdf of U and Z.

For the upper bound, we utilize the classic normal tail bound, $\mathbb{P}(x < Z) \leqslant e^{-x^2/2}/(x\sqrt{2\pi})$.

$$P \leqslant \mathbb{E}\left[\frac{e^{-\left(\delta+\sqrt{U}\sqrt{k^{2/p}-1}\right)^2/2}}{(\delta+\sqrt{U}\sqrt{k^{2/p}-1})\sqrt{2\pi}}\right] \leqslant \frac{e^{-\delta^2/2}}{\delta\sqrt{2\pi}} \mathbb{E}\left[e^{-\delta\sqrt{U}\sqrt{k^{2/p}-1}-\frac{1}{2}U\left(k^{2/p}-1\right)}\right]. \tag{37}$$

Observing the bound that for positive x, $e^{-x} \leqslant 1/x$, and recalling that $d \geqslant 2$,

$$\begin{aligned} P &\leqslant \frac{e^{-\delta^2/2}}{\delta\sqrt{2\pi}} \mathbb{E}\left[\frac{e^{-\frac{1}{2}U\left(k^{2/p}-1\right)}}{\delta\sqrt{U}\sqrt{k^{2/p}-1}}\right] \\ &= \frac{e^{-\delta^2/2}}{\delta^2\sqrt{2\pi}\sqrt{k^{2/p}-1}} \mathbb{E}\left[U^{-\frac{1}{2}} e^{-\frac{1}{2}U\left(k^{2/p}-1\right)}\right] \\ &= \frac{e^{-\delta^2/2}}{\delta^2\sqrt{2\pi}\sqrt{k^{2/p}-1}} \left(\frac{k^{(1-d)/p}\Gamma\left(\frac{d}{2}-\frac{1}{2}\right)}{\sqrt{2}\Gamma\left(\frac{d}{2}\right)}\right). \end{aligned} \tag{38}$$

Here we utilize the following bounds: $e^x - 1 \geqslant (e/2)x^2$, which is easy to prove, and $\Gamma(d/2-1/2)/\Gamma(d/2) \leqslant \sqrt{2\pi/d}$, which may be proved on integer $d \geqslant 2$ by induction. This yields:

$$P \leqslant \frac{e^{-(1+\delta^2)/2} p}{2\delta^2 \ln k} \frac{k^{(1-d)/p}}{\sqrt{d}}. \tag{39}$$

This completes the proof.

Remark 4 Room for Improvement: The choice of the $e^x - 1 \geqslant (e/2)x^2$ bound above was in fact arbitrary - other bounds, such as involving alternative powers of x, could be used. This would influence how the resulting bound on P is utilized, for instance in the proof of Theorem 3. The use of $e^{-x} \leqslant 1/x$ in Eq. (38) should be considered similarly. ■

Proposition 6 *Conjecture 1 is* **false** *and for each i, for* $\varepsilon > 0$*,*

$$\frac{\mathbb{P}\left(\bar{X}^i_j + S_i(j)\sqrt{k^{2/j}-1} < \mu_i - \varepsilon \textit{ for some } 2 \leqslant j \leqslant k\right)}{1/k} \to \infty \textit{ as } k \to \infty. \tag{40}$$

.

Proof [of Proposition 6] Define the events $A^i_{j,k,\varepsilon} = \{\bar{X}^i_j + S_i(j)\sqrt{k^{2/j}-1} < \mu_i - \varepsilon\}$. As the samples are taken to be normally distributed with mean μ_i and variance σ_i^2, we have that $\bar{X}^i_j - \mu_i \sim Z\sigma_i/\sqrt{j}$ and $S_i^2(j) \sim \sigma_i^2 U/j$,

where Z is a standard normal, $U \sim \chi^2_{j-1}$, and Z, U independent. Hence,

$$\mathbb{P}(A^i_{j,k,\varepsilon}) = \mathbb{P}\left(Z\frac{\sigma_i}{\sqrt{j}} + \sqrt{U\frac{\sigma_i^2}{j}}\sqrt{k^{2/j}-1} < -\varepsilon\right) = \mathbb{P}\left(\frac{\varepsilon}{\sigma_i}\sqrt{j} + \sqrt{U}\sqrt{k^{2/j}-1} < Z\right). \tag{41}$$

The last step is simply a re-arrangement, and an observation on the symmetry of the distribution of Z. For $j \geqslant 3$, we may apply Proposition 2 here for $d = j-1$, $p = j$, to yield

$$\mathbb{P}(A^i_{j,k,\varepsilon}) \geqslant \frac{1}{2}\frac{k^{1/j}}{k}\mathbb{P}\left(\frac{1}{4}Z^2 \geqslant U \geqslant \frac{\varepsilon^2}{\sigma_i^2}j\right). \tag{42}$$

For a fixed $j_0 \geqslant 3$, for $k \geqslant j_0$ we have

$$\mathbb{P}\left(A^i_{j,k,\varepsilon} \text{ for some } 2 \leqslant j \leqslant k\right) \geqslant \mathbb{P}(A^i_{j_0,k,\varepsilon}) \geqslant O(1/k)\, k^{1/j_0}. \tag{43}$$

The proposition follows immediately. ■

Proposition 7 *For $G > 0$, $0 \leqslant \varepsilon < 1/2$, the following holds:*

$$\frac{1}{\ln\left(1 + G\frac{(1-\varepsilon)^2}{1+\varepsilon}\right)} \leqslant \frac{1}{\ln(1+G)} + \frac{10G}{(1+G)\ln(1+G)^2}\varepsilon. \tag{44}$$

Proof For any $G > 0$, the function $1/\ln\left(1 + G\frac{(1-\varepsilon)^2}{1+\varepsilon}\right)$ is positive, increasing, and convex on $\varepsilon \in [0,1)$ (Proposition 8). For a given $G > 0$, noting that the above inequality holds (as equality) at $\varepsilon = 0$, due to the convexity it suffices to show that the inequality is satisfied at $\varepsilon = 1/2$, or

$$\frac{1}{\ln\left(1 + \frac{G}{6}\right)} \leqslant \frac{5G}{(1+G)\ln(1+G)^2} + \frac{1}{\ln(1+G)}. \tag{45}$$

Equivalently, we consider the inequality

$$0 \leqslant \frac{5G}{(1+G)} + \ln(1+G) - \frac{\ln(1+G)^2}{\ln\left(1+\frac{G}{6}\right)}. \tag{46}$$

Define the function $F(G)$ to be the RHS of Ineq. (46). Note that as $G \to 0$, $F(G) \to 0$, and in simplified form we have (for $G > 0$ and the limit as $G \to 0$),

$$F'(G) = \frac{\left((1+G)\ln(1+G) - (6+G)\ln\left(1+\frac{G}{6}\right)\right)^2}{(1+G)^2(6+G)\ln\left(1+\frac{G}{6}\right)^2} \geqslant 0. \tag{47}$$

It follows that $F(G) \geqslant 0$, and hence the desired inequality holds at $\varepsilon = 1/2$. This completes the proof. ■

Proposition 8 *The function $H_G(\varepsilon) = 1/\ln\left(1 + G\frac{(1-\varepsilon)^2}{1+\varepsilon}\right)$ is positive, increasing, and convex in $\varepsilon \in [0,1)$, for any constant $G > 0$.*

Proof That $H_G(\varepsilon)$ is positive and increasing in ε, follows immediately from inspection of H_G and H'_G, given the hypotheses on G, and ε. To demonstrate convexity, by inspection of the terms of $H''_G(\varepsilon)$, it suffices to show that for all relevant G, and ε, the following inequality holds.

$$2G(1-\varepsilon)^2(3+\varepsilon)^2 + \big(-8(1+\varepsilon) + G(1-\varepsilon)^2(1+\varepsilon(6+\varepsilon))\big)\ln\left(1+G\frac{(1-\varepsilon)^2}{1+\varepsilon}\right) \geqslant 0. \tag{48}$$

Defining $C = G(1-\varepsilon)^2/(1+\varepsilon)$, it is sufficient to show that for all $C > 0$ and $\varepsilon \in [0,1)$ (eliminating a factor of $(1+\varepsilon)$ from the above),

$$2C(3+\varepsilon)^2 + (-8 + C(1+\varepsilon(6+\varepsilon)))\ln(1+C) \geqslant 0. \tag{49}$$

Defining $J_C(\varepsilon)$ as the LHS of the above, note that $J'_C(\varepsilon) = 2C(3+\varepsilon)(2+\ln(1+C)) > 0$. It suffices then to show $J_C(0) \geqslant 0$, or $18C + (C-8)\ln(1+C) \geqslant 0$. Note this holds at $C = 0$, and $d/dC[J_C(0)] = (10+19C)/(1+C) + \ln(1+C) > 0$ for $C \geqslant 0$. Hence, $J_C(\varepsilon) \geqslant 0$, and $H''_G(\varepsilon) \geqslant 0$. ■